The Self-Organized Archive: SPASE, PDS and Archive Cooperatives

Todd A. King[1], J. Steven Hughes[2], D. Aaron Roberts[3], Raymond J. Walker[1], and Steven P. Joy[1]

[1]Institute of Geophysics and Planetary Physics, UCLA, 3846 Slichter Hall, Los Angeles, CA 90095-1567 USA

[2]Jet Propulsion Laboratory, Pasadena, CA; [3]NASA Goddard Space Flight Center, Greenbelt, MD
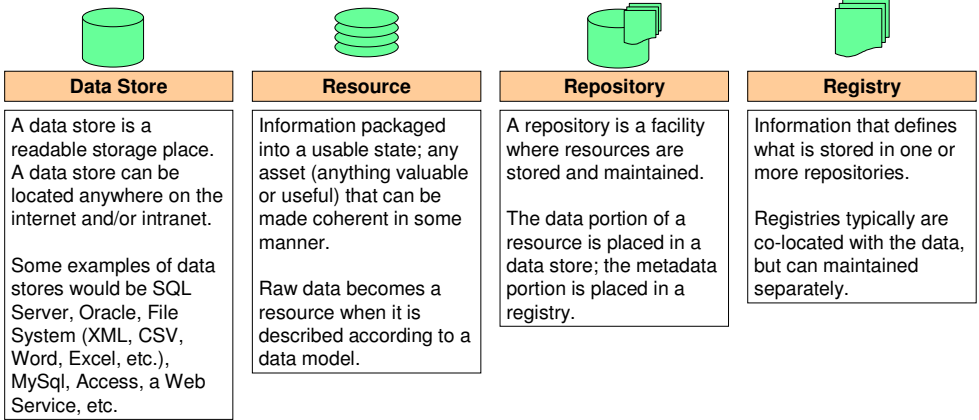
UCLA

What is a Self-Organized Archive

Any data store with sufficient metadata to permit incremental additions and a search capability to locate all related items.

-or-

A repository that permits incremental additions and capability to locate all related items.

Data System Components

Data Store

A data store is a readable storage place. A data store can be located anywhere on the internet and/or intranet.

Some examples of data stores would be SQL Server, Oracle, File System (XML, CSV, Word, Excel, etc.), MySql, Access, a Web Service, etc.

Resource

Information packaged into a usable state; any asset (anything valuable or useful) that can be made coherent in some manner.

The data portion of a resource is placed in a data store; the metadata portion is placed in a registry.

Raw data becomes a resource when it is described according to a data model.

Repository

A repository is a facility where resources are stored and maintained.

The data portion of a resource is placed in a data store; the metadata portion is placed in a registry.

Registry

Information that defines what is stored in one or more repositories.

Registries typically are co-located with the data, but can maintained separately.

Types of Archives

There are many types and architectures for archives.

Stack Archive: Resources are loaded into the repository from multiple sources. (i.e. a single library)

Distributed Stack: Resources are loaded into "local" repositories; registry information is shared. (i.e., branch libraries, PDS)

Trusted Peer-to-peer: Resources are stored at provider sites; registry information is shared; membership is required. (i.e., VSPO)
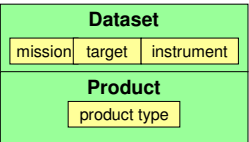
Open Peer-to-peer: Resources are stored at provider sites; registry information is shared. (i.e., Napster, Kazaa, and other things to get in trouble with)

What is Needed to Enable Self-Organization

A data model with sufficient abstraction so that different types of resources can declare membership in an abstract concept.
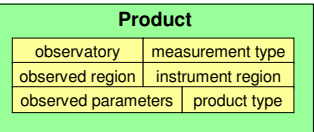
PDS

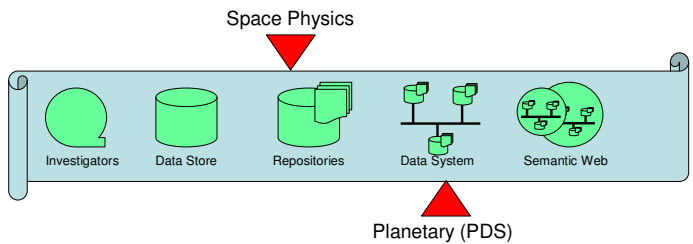The PDS data model has the following abstract layers:

Dataset		
mission	target	instrument
Product		
product type		

SPASE

The SPASE data model has the following abstract layers:

Product		
observatory	measurement type	
observed region	instrument region	
observed parameters	product type	

System Model Evolution

Space Physics

Planetary (PDS)

The distribution of data has evolved from tightly coupled networks of people with partially coupled data to tightly coupled data (repositories) with partially coupled networks of people. The next step in the evolution is lightly coupled, self-organizing networks of data and partially coupled networks of people and systems (the Semantic Web).

How Does an Archive Self-Organize

A self-organizing archive discovers the association of resources during the search process. Such associations are not inherent in the organization of the archive.

Some examples:

1. **A PDS volume** embeds associative metadata in the file system. Documents go in a document directory, data in a data directory. Presumably only documents that relate to the data are stored on the volume. This determination is done by a data engineer.

   However, if additional metadata is associated with the documents, then a volume can self-organize in response to a query.

2. **A researcher makes a discovery.** The observations and interpretation are placed in a document and the appropriate metadata is registered in the system. Another researcher locates resources used in the discovery and looks for related resources. The researcher's discovery is then revealed to the other researcher.

Benefits of a Self-Organizing Archive

- Resource providers are only concerned with the resources they provide.
- Providers document their resource to published standards. Everyone's efforts are complimentary and easily leveraged.
- No need for "build" procedures. Low management overhead.
- Better match with reality (resources are added as they become available).
- Associations are "organic". What belongs together is brought together.

The Semantic Web

- Semantic technology uses ontologies (formal specifications useful in conceptualization) to discover the relevance of resources in a single domain and between multiple domains.
  - An ontology defines the relationships between objects defined in the data model.
  - Formally mapping these relationships gives meaning to the metadata. For example,
    - ☐ Are missions and observatories equivalent?
    - ☐ Are "different" instruments, like cosmic ray detectors and energetic particle detectors, equivalent?

How Much Semantic Modeling Do You Need?

- Some semantic features can be handled in the data model through abstraction. This usually applies only within a single domain. You don't always need a semantic layer.
- Domain-to-domain relevance (for example between earth science and space physics) requires semantic technology to resolve issues of meaning.

Summary

- Developing self-organizing archives is the next step in data system evolution. It can be achieved by building trusted peer-to-peer data systems. Associations or relevance of one resource to another can be determined through well defined and enforced data models.
- Self-organized archives can be implemented by adopting semantic technologies, but within a single domain (data model) the benefit is limited and difficult to justify.
- Broad, multi-domain, self-organized archives require semantic technologies, especially when coordination between domains is difficult and each domain has existing and well established data models.

For More Information

Todd King
IGPP/UCLA
3846 Slichter Hall
Los Angeles, CA 90095-1567

(310) 206-7201

tking@igpp.ucla.edu